Severe infections emerge from commensal bacteria by adaptive evolution

Article (Published Version)

eLIFE elifesciences.org

RESEARCH ARTICLE

Severe infections emerge from commensal bacteria by adaptive evolution

Bernadette C Young[1,2]*, Chieh-Hsi Wu[1], N Claire Gordon[1], Kevin Cole[3], James R Price[3,4], Elian Liu[1,2], Anna E Sheppard[1,5], Sanuki Perera[1,2], Jane Charlesworth[1], Tanya Golubchik[1], Zamin Iqbal[6], Rory Bowden[6], Ruth C Massey[7], John Paul[8,9], Derrick W Crook[1,8,9], Timothy E Peto[1,9], A Sarah Walker[1,9], Martin J Llewelyn[3,4], David H Wyllie[1,10], Daniel J Wilson[1,6,11]*

[1]Nuffield Department of Medicine, Experimental Medicine Division, University of Oxford, Oxford, United Kingdom; [2]Microbiology and Infectious Diseases Department, Oxford University Hospitals NHS Foundation Trust, Oxford, United Kingdom; [3]Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton, United Kingdom; [4]Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Brighton, United Kingdom; [5]NIHR Health Protection Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, United Kingdom; [6]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom; [7]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom; [8]National Infection Service, Public Health England, London, United Kingdom; [9]National Institute for Health Research, Oxford Biomedical Research Centre, Oxford, United Kingdom; [10]Centre for Molecular and Cellular Physiology, Jenner Institute, Oxford, United Kingdom; [11]Institute for Emerging Infections, Oxford Martin School, University of Oxford, Oxford, United Kingdom

***For correspondence:** bernadette.young@ndm.ox.ac.uk (BCY); daniel.wilson@ndm.ox.ac.uk (DJW)

Competing interests: The authors declare that no competing interests exist.

Funding: See page 16

Received: 21 July 2017
Accepted: 02 December 2017
Published: 19 December 2017

Reviewing editor: Matthew TG Holden, University of St Andrews, United Kingdom

Abstract Bacteria responsible for the greatest global mortality colonize the human microbiota far more frequently than they cause severe infections. Whether mutation and selection among commensal bacteria are associated with infection is unknown. We investigated de novo mutation in 1163 *Staphylococcus aureus* genomes from 105 infected patients with nose colonization. We report that 72% of infections emerged from the nose, with infecting and nose-colonizing bacteria showing parallel adaptive differences. We found 2.8-to-3.6-fold adaptive enrichments of protein-altering variants in genes responding to *rsp*, which regulates surface antigens and toxin production; *agr*, which regulates quorum-sensing, toxin production and abscess formation; and host-derived antimicrobial peptides. Adaptive mutations in pathogenesis-associated genes were 3.1-fold enriched in infecting but not nose-colonizing bacteria. None of these signatures were observed in healthy carriers nor at the species-level, suggesting infection-associated, short-term, within-host selection pressures. Our results show that signatures of spontaneous adaptive evolution are specifically associated with infection, raising new possibilities for diagnosis and treatment.
DOI: https://doi.org/10.7554/eLife.30637.001

Introduction

Infections remain a leading cause of global mortality, with bacterial pathogens among the greatest concern (***GBD 2015 Mortality and Causes of Death Collaborators, 2016***). However, many of the bacteria imposing the greatest burden of mortality, such as *Staphylococcus aureus*, are frequently

found as commensal components of the body's microbiota (**Turnbaugh et al., 2007**). For them, infection is a relatively uncommon event that is often unnecessary (**Casadevall et al., 2011**; **Méthot and Alizon, 2014**), and perhaps disadvantageous (**Brown et al., 2012**), for onward transmission. Genomics is shedding light on important bacterial traits such as host-specificity, toxicity and antimicrobial resistance (**Sheppard et al., 2013**; **Laabei et al., 2014**; **Chewapreecha et al., 2014**; **Chen et al., 2015**; **Earle et al., 2016**). These approaches offer new opportunities to understand the role of genetics and within-host evolution in the outcome of human interactions with major bacterial pathogens (**Didelot et al., 2016**).

Several lines of evidence support a plausible role for within-host evolution influencing the virulence of bacterial pathogens. Common bacterial infections, including *S. aureus*, are often associated with colonization of the nose by a genetically similar strain. In these patients, the nose is considered the likely source of infection because the nose is more often the site of asymptomatic colonization than any other body site (**von Eiff et al., 2001**; **Kluytmans et al., 1997**; **Yang et al., 2010**). Genome sequencing suggests that bacteria mutate much more quickly than previously accepted, and this confers a potent ability to adapt, for example evolving antimicrobial resistance de novo within individual patients (**Howden et al., 2011**; **Eldholm et al., 2014**). Opportunistic pathogens infecting cystic fibrosis patients have been found to rapidly adapt to the lung environment, with strong evidence of parallel evolution across patients (**Lieberman et al., 2011**; **Marvig et al., 2013**; **Markussen et al., 2014**; **Lieberman et al., 2014**; **Marvig et al., 2015**). However, the selection pressures associated with antimicrobial resistance and opportunistic infections of cystic fibrosis patients may not typify within-host adaptation in common commensal pathogens that have co-evolved with humans for thousands or millions of years (**Moeller et al., 2016**; **Lees et al., 2017**).

Candidate gene studies have demonstrated that substitutions in certain regions, notably quorum-sensing systems such as the *S. aureus* accessory gene regulator (*agr*), arise particularly quickly in vivo and in culture (**Traber et al., 2008**). The *agr* operon encodes a pheromone that coordinates a shift at higher cell densities from production of surface proteins promoting biofilm formation to production of secreted toxins and proteases promoting inflammation and dispersal (**Novick and Geisinger, 2008**). Mutants typically produce the pheromone but no longer respond to it (**Painter et al., 2014**). The evolution of *agr* has been variously ascribed to directional selection (**Sakoulas et al., 2009**), balancing selection (**Robinson et al., 2005**), social cheating (**Pollitt et al., 2014**) and life-history trade-off (**Shopsin et al., 2010**). However, the role of *agr* mutants in infection remains unclear, since they are frequently sampled from both asymptomatic carriage and severe infections (**Smyth et al., 2012**; **Painter et al., 2014**).

Whole-genome sequencing case studies add weight to the idea that within-host evolution plays an important role in infection. In one persistent *S. aureus* infection, a single mutation was sufficient to permanently activate the stringent stress response, reducing growth, colony size and experimentally measured infection severity (**Gao et al., 2010**). In another patient, we found that bloodstream bacteria differed from those initially colonizing the nose by several mutations including loss-of-function of the *rsp* regulator (**Young et al., 2012**). Functional follow-up revealed that the *rsp* mutant expressed reduced cytotoxicity (**Laabei et al., 2015**), but maintained the ability to cause disseminated infection (**Das et al., 2016**). Unexpectedly, we found that bloodstream-infecting bacteria exhibit lower cytotoxicity than nose-colonizing bacteria more generally (**Laabei et al., 2015**). These results raise the question: are unique hallmarks of de novo mutation and selection associated with bacterial evolution in severely infected patients?

We addressed this question by investigating the genetic variants arising from within-patient evolution of *S. aureus* sampled from 105 patients with concurrent nose colonization and blood or deep tissue infection. We annotated variants to test for systematic differences between colonizing and infecting bacteria. We discovered several groups of genes showing significant enrichments of protein-altering variants compared to other genes, indicating adaptive evolution. For genes implicated in pathogenesis, adaptive mutants were limited to infecting bacteria, while other pathways showed adaptation in the nose and infection site. Adaptive enrichments were not observed in asymptomatic carriers, nor between unrelated bacteria, indicating evolution in response to infection-associated, within-host selection pressures. Our results reveal that adaptive evolution of genes involved in regulation, toxin production, abscess formation, cell-cell communication and bacterial-host interaction drives parallel differentiation between commensal constituents of the nose and infecting bacteria, providing new insights into the evolution of infection in a major pathogen.

Results

Infecting bacteria are typically descended from the patient's commensal bacteria

We identified 105 patients suffering severe *S. aureus* infections admitted to hospitals in Oxford and Brighton, England, for whom we could recover contemporaneous nose swabs from admission screening. Of the 105 patients, 55 had bloodstream infections, 37 had soft tissue infections and 13 had bone and joint infections (***Table 1***). The infection was most often sampled on the same day as the nose, with an interquartile range of 1 day earlier to 2 days later (***Supplementary file 1***).

To discover de novo mutations within and between *S. aureus* in the nose and infection site, we whole-genome sequenced 1163 bacterial colonies, a median of 5 per site. We detected single-nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) using previously developed, combined reference-based mapping and de novo assembly approaches (***Young et al., 2012***; ***Golubchik et al., 2013***; ***Iqbal et al., 2012***). We identified 35 distinct strains, defined by multilocus sequence type (ST), across patients (***Supplementary file 1***). As expected (***von Eiff et al., 2001***), most patients possessed extremely closely related nose-colonizing and infecting bacteria, sharing the same ST and differing by 0–66 variants (95 patients). The nose-colonizing and infecting bacteria of nine patients were unrelated, possessing different STs and differing by 9398–50573 variants (e.g. ***Figure 1A***). In one further patient, we deemed the nose-colonizing and infecting bacteria to be unrelated despite sharing the same ST because they differed by 1104 variants, far outside the within-ST variation evident in any individual nose or infection site (***Figure 1—figure supplement 1***), and corresponding to around 70 years of divergence based on our previous estimates of within-host evolution (***Young et al., 2012***). In 9/95 patients with extremely closely related nose-colonizing and infecting bacteria, another, unrelated ST was also present in the nose (six patients) or the infection site (three patients); we excluded these unrelated bacteria from further analysis. After excluding variants differentiating unrelated nose-colonizing and infecting bacteria, we catalogued 1322 de novo mutations that we deemed arose within the 105 patients.

In patients with closely related strains, the within-patient population structure was always consistent with a unique migration event from the nose to the infection site, or occasionally, vice versa. Infecting and nose-colonizing bacteria usually formed closely related but distinct populations with no shared genotypes (74/95 patients, e.g. ***Figure 1B***), separated by a mean of 5.7 variants. There was never more than one identical genotype between nose-colonizing and infecting bacteria, (21/95 patients, e.g. ***Figure 1C***), indicating that the migration event from one population to the other involved a small number of founding bacteria (***Moxon and Murphy, 1978***; ***Margolis and Levin, 2007***; ***Prajsnar et al., 2012***). In such patients, the shared genotype likely represents the migrating genotype itself. Population structure did not differ significantly between infection types (p=0.38, ***Table 1***). Genetic diversity in the nose (mean pairwise distance, π = 2.8 variants) was similar to that previously observed in asymptomatic nasal carriers (***Golubchik et al., 2013***) (Reference Panel I, π = 4.1, p=0.13), but was significantly lower in the infection site (π = 0.6, p=$10^{-10.0}$), revealing limited diversification post-infection.

In most patients, the infection appeared to be descended from the nose. We used 1149 sequences from other patients and carriers (Reference Panel II) to reconstruct the most recent common

Table 1. Distribution of infection types and relatedness of nose-colonizing and infecting *S. aureus* among 105 patients revealed by genomic comparison.

	Relation of nose-colonizing to infecting bacteria		
		Closely related (≤66 variants)	
Infection sites	Unrelated (≥1104 variants)	Zero shared genotypes	One shared genotype
Bloodstream	4	43	8
Soft tissue	4	23	10
Bone and joint	2	8	3
Total	10	74	21

DOI: https://doi.org/10.7554/eLife.30637.002

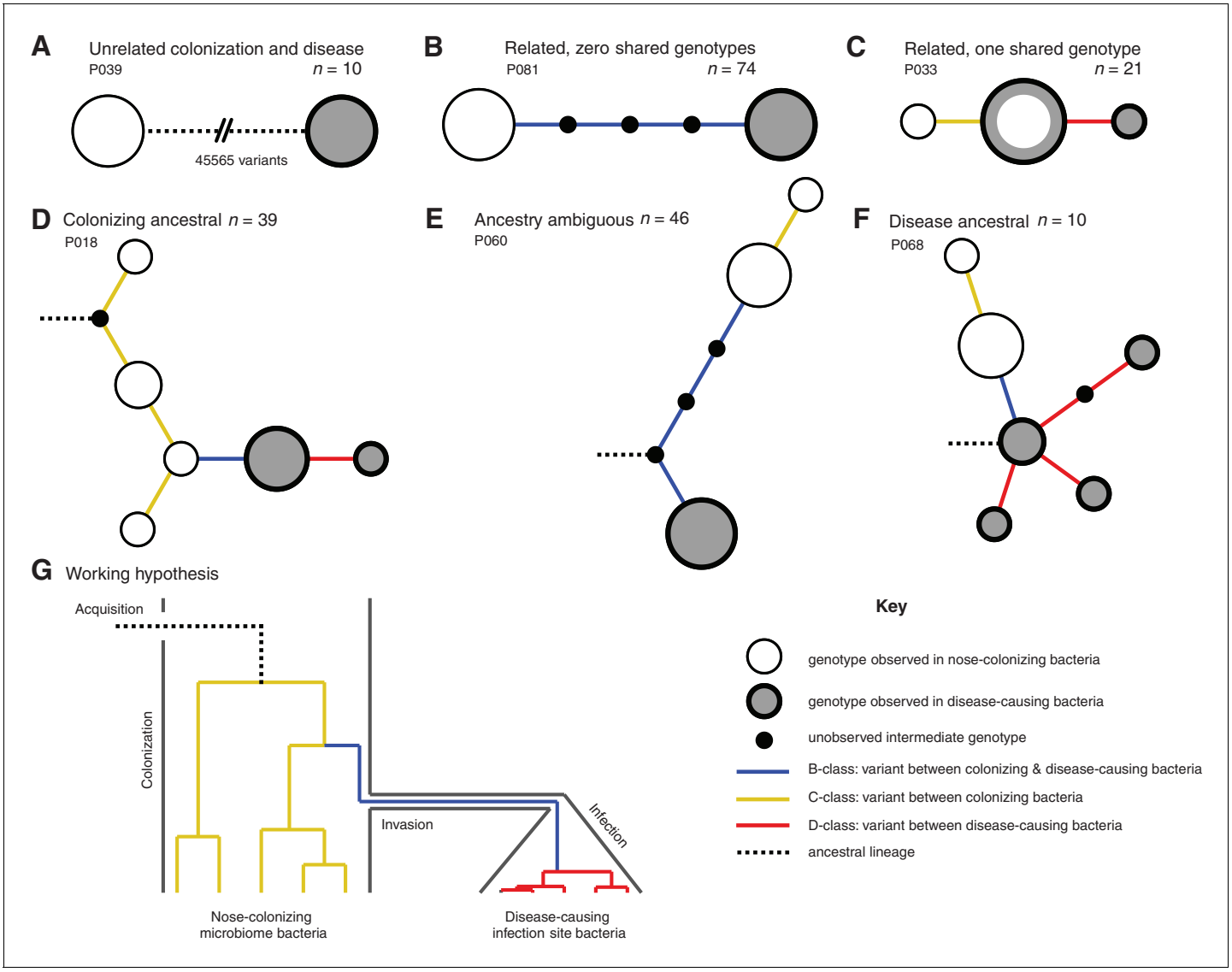

Figure 1. Infection-causing *S. aureus* form closely related but distinct populations descended from nose-colonizing bacteria in the majority of infections. Bacteria sampled from the nose and infection site of 105 patients formed one of three population structures, illustrated with example haplotrees: (**A**) Unrelated populations differentiated by many variants. (**B**) Highly related populations separated by few variants. (**C**) Highly related populations with one genotype in common. Reconstructing the ancestral genotype in each patient helped identify the ancestral population: (**D**) Nose-colonizing bacteria ancestral. (**E**) Ambiguous ancestral population. (**F**) Infection site bacteria ancestral. (**G**) Phylogeny illustrating the working hypothesis that variants differentiating highly related nose-colonizing and infection-causing bacteria would be enriched for variants that promote, or are promoted by, infection. In **A–F**, haplotree nodes represent observed genotypes sampled from the nose (white) or infection site (grey), with area proportional to genotype frequency, or unobserved intermediate genotypes (black). Edges represent mutations. Patient identifiers and sample sizes (**n**) are given. In **A–G**, edge color indicates that mutations occurring on those branches correspond to B-class variants between nose-colonizing and infection-causing bacteria (blue), C-class variants among nose-colonizing bacteria (gold) or D-class variants among infection-causing bacteria (red). Black dashed edges indicate ancestral lineages. A B C.

DOI: https://doi.org/10.7554/eLife.30637.003

The following figure supplement is available for figure 1:

Figure supplement 1. Distribution of the number of variants identified within 105 severely infected patients, by class.

DOI: https://doi.org/10.7554/eLife.30637.004

ancestor (MRCA) for the 95/105 (90%) patients with related nose-colonizing and infecting bacteria. We thereby distinguished wild type from mutant alleles. In 49 such patients, we could determine the ancestral population. The nose population was likely ancestral in 39/49 (80% of patients with related strains, or 72% of all patients) because all infecting bacteria shared de novo mutations in common that distinguished them from the MRCA, whereas nose-colonizing bacteria did not. In 16 of those, confidence was high because both mutant and ancestral alleles were observed in the nose, confirming it as the origin of the de novo mutation (e.g. ***Figure 1D***). Conversely, in 10/49 patients, bacteria colonizing the nose were likely descended from blood or deep tissue infections (20% of patients with related strains, or 18% of all patients) (e.g. ***Figure 1F***). Confidence was high for just three of those patients, and they showed unusually high diversity (Supplementary data, P063, P072, P093), suggesting that in persistent infections, infecting bacteria can recolonize the nose.

Protein-truncating mutants are over-represented within infected patients

To help identify variants that could promote, or be promoted by, infection of the blood and deep tissue by bacteria colonizing the nose, we reconstructed within-patient phylogenies and classified variants by their position in the phylogeny. Sequencing multiple colonies per site enabled us to classify variants into those representing genuine differences *between* nose-colonizing and infection populations (*B*-class), variants specific to the nose-*colonizing* population (*C*-class) and variants specific to the *disease*-causing infection population (*D*-class). We hypothesized that B-class variants would be most enriched for variants promoting, or promoted by, infection, if such variants occur (***Figure 1G***).

We cross-classified variants by their predicted functional effect: synonymous, non-synonymous or truncating within protein-coding sequences, or non-coding (***Table 2***, ***Supplementary file 2***). As expected, the prevailing tendency of selection within patients was to conserve protein sequences, with d_N/d_S ratios indicating rates of non-synonymous change 0.55, 0.68 and 0.63 times the rate expected under strict neutral evolution for B-, C- and D-class variants, respectively.

In a longitudinal study of one long-term carrier, we previously reported that a burst of protein-truncating variants punctuated the transition from asymptomatic nose carriage to infection (***Young et al., 2012***). Here, we found a 3.9-fold over-abundance of protein-truncating variants of all phylogenetic classes in infected patients compared to asymptomatic carriers (Reference Panel I, p=0.002, ***Table 2***), supporting the conclusion that loss-of-function mutations are disproportionately associated with evolution within infected patients. This may reflect a reduction in the efficiency with which selection removes deleterious protein-truncating mutations, and provides evidence of a systematic difference in selection within severely infected patients.

Table 2. Cross-classification of variants within patients by phylogenetic position and predicted functional effect, and comparison to asymptomatic nose carriers.

Neutrality indices (***McDonald and Kreitman, 1991***; ***Rand and Kann, 1996***) were defined as the odds ratio of mutation counts relative to synonymous variants in patients versus asymptomatic nose carriers (Reference Panel I). Those significant at p<0.05 and p<0.005 are emboldened and underlined respectively.

	Number of variants (Neutrality index)				
Phylogenetic position	**Synonymous**	**Non-synonymous**	**Protein truncating**	**Non-coding**	**Total**
Patients with severe infections (*n* = 105)					
Between nose-colonization and infection site (B-class)	93	265 (1.1)	**39 (3.1)**	140 (1.2)	537
Within nose-colonization (C-class)	93	325 (1.3)	**59 (4.7)**	145 (1.3)	622
Within infection site (D-class)	26	82 (1.2)	**15 (4.3)**	40 (1.3)	163
Total	212	672 (1.2)	**113 (3.9)**	325 (1.3)	1322
Asymptomatic carriers (***Golubchik et al., 2013***) (Reference panel I, for comparison, *n* = 13)					
Within nose-colonization (C-class)	37	97	5	45	184

DOI: https://doi.org/10.7554/eLife.30637.005

Quorum sensing and cell-adhesion proteins exhibit adaptive evolution between nose-colonizing and infecting bacteria

We hypothesized that variants associated with infection would be enriched among the protein-altering B-class variants between the nose and infection site (***Figure 1G***). Therefore, we aggregated mutations by genes in a well-annotated reference genome, MRSA252, and tested each gene for an excess of non-synonymous and protein-truncating B-class variants compared to other genes, taking into account the length of the genes. Aggregating by gene was necessary because 1318/1322 variants were unique to single patients. The two exceptions involved non-coding variants arising in two patients each, one B-class variant 130 bases upstream of *azlC*, an azaleucine resistance protein (SAR0010), and one D-class variant 88 bases upstream of *eapH1*, a secreted serine protease inhibitor (SAR2295) (***Stapels et al., 2014***).

We found a significant excess of five protein-altering B-class variants representing a 58.3-fold enrichment in *agrA*, which encodes the response regulator that mediates activation of the quorum-sensing system at high cell densities ($p=10^{-7.5}$, ***Figure 2A***, ***Table 3***). The *clfB* gene encoding clumping factor B, which binds human fibrinogen and loricrin (***Foster et al., 2013***), showed an excess of five protein-altering B-class variants, representing a 15.9-fold enrichment that was near genome-wide significance after multiple testing correction ($p=10^{-4.7}$). Both signals of enrichment produced neutrality indices exceeding one, consistent with adaptive evolution (***Supplementary file 3***).

Previously, we identified a truncating mutation in the transcriptional regulator *rsp* to be the most likely candidate for involvement in the progression to infection in one long-term nasal carrier (***Young et al., 2012***). Although we observed just one variant in *rsp* among the 105 patients (3.9-fold enrichment, p=0.27), we found it was a non-synonymous B-class variant resulting in an alanine to proline substitution in the regulator's helix-turn-helix DNA binding domain. In separately published

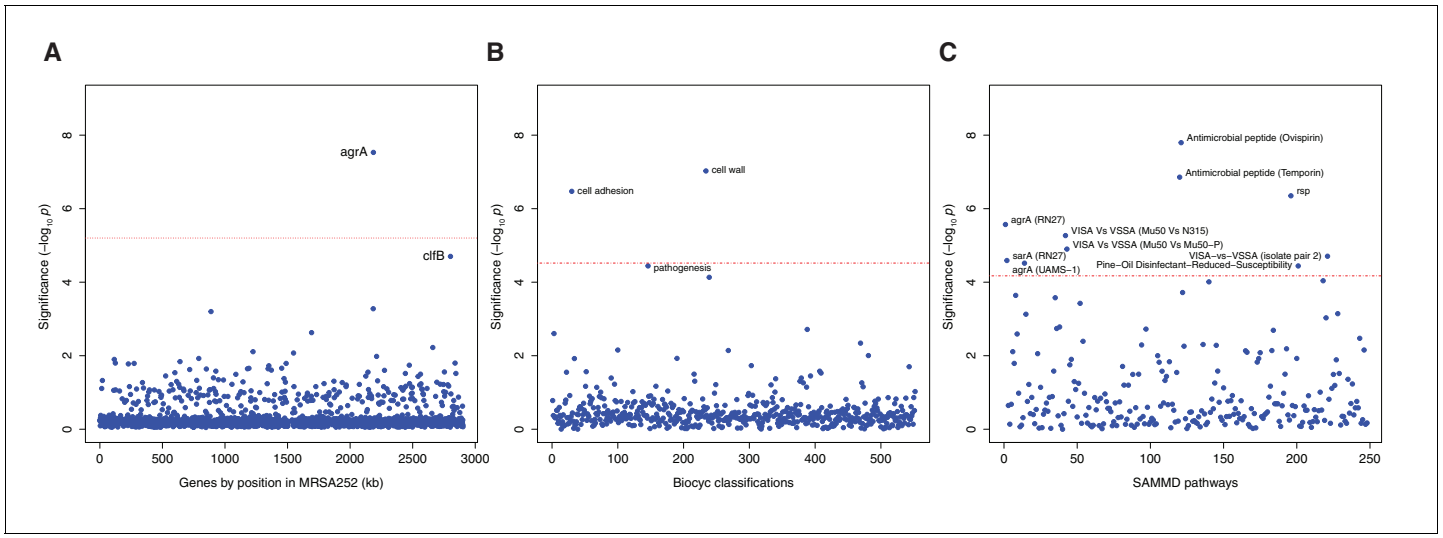

Figure 2. Genes, ontologies and pathways enriched for protein-altering substitutions between nose-colonizing and infection-causing bacteria within infected patients. (**A**) Significance of enrichment of 2650 individual genes. (**B**) Significance of enrichment of 552 gene sets defined by BioCyc gene ontologies. (**C**) Significance of enrichment of 248 gene sets defined by SAMMD expression pathways. Genes, pathways and ontologies that approach or exceed a Bonferroni-corrected significance threshold of $\alpha = 0.05$, weighted for the number of tests per category, (red lines) are named.

DOI: https://doi.org/10.7554/eLife.30637.006

The following figure supplements are available for figure 2:

Figure supplement 1. Genes, ontologies and pathways enriched for protein-altering transient variants within nose-colonizing and infection-causing bacteria.

DOI: https://doi.org/10.7554/eLife.30637.007

Figure supplement 2. Gene set enrichment analysis of B-class mutants occurring in the nose or the infection site.

DOI: https://doi.org/10.7554/eLife.30637.008

Figure supplement 3. Genes, ontologies and pathways enriched for protein-altering variants among longitudinally sampled asymptomatic nasal carriers.

DOI: https://doi.org/10.7554/eLife.30637.009

Table 3. Genes, gene ontologies and expression pathways exhibiting the most significant enrichments or depletions of protein-altering B-class variants separating nose and infection site bacteria.

Enrichments below one represent depletions. The total number of variants and genes available for analysis differed by database. A -$\log_{10}$ p-value above 5.2, 4.5 or 4.2 was considered genome-wide significant for loci, gene ontologies or expression pathways respectively (in bold).

Gene group	No. protein-altering B-class variants	Cumulative lengthof genes (kb)	Enrichment	Significance (-$\log_{10}$ p value)
Locus				
agrA	5	0.7	58.27	**7.53**
clfB	5	2.6	15.87	4.70
Total	289	2363.8		
BioCyc Gene Ontology (***Caspi et al., 2016***)				
Cell wall	18	30.9	5.02	**7.03**
Cell adhesion	13	17.2	6.44	**6.47**
Pathogenesis	31	112.5	2.41	4.44
Total	288	2359.3		

SAMMD Expression Pathway	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	*Down-regulated*	*Up-regulated*	
Ovispirin-1 (***Pietiäinen et al., 2009***)	40	7	121.2	142.9	2.65	0.39	**7.80**
Temporin L (***Pietiäinen et al., 2009***)	42	14	125.1	156.1	2.78	0.74	**6.86**
rsp (***Lei et al., 2011***)	27	1	61.1	13.7	3.61	0.60	**6.35**
agrA (RN27) (***Dunman et al., 2001***)	9	30	41.0	85.0	1.83	2.94	**5.57**
VISA-vs-VSSA (Mu50 vs N315) (***Cui et al., 2005***)	0	17	0	34.4	0	3.95	**5.27**
VISA-vs-VSSA (Mu50 vs Mu50-P) (***Cui et al., 2005***)	0	17	0	36.7	0	3.70	**4.90**
VISA-vs-VSSA (isolate pair 2) (***Howden et al., 2008***)	14	3	26.9	59.7	4.06	0.39	**4.71**
sarA (RN27) (***Dunman et al., 2001***)	6	23	49.9	57.7	0.97	3.22	**4.59**
agrA (UAMS-1 OD 1.0) (***Cassat et al., 2006***)	0	5	0	2.7	0	14.57	**4.52**
Pine-Oil Disinfectant-Reduced-Susceptibility (***Lamichhane-Khadka et al., 2008***)	17	5	36.4	23.6	3.76	1.70	**4.44**
Total	275		2093.5				

DOI: https://doi.org/10.7554/eLife.30637.010

experiments (***Das et al., 2016***), we demonstrated that this and the original mutation induce similar loss-of-function phenotypes which, like *agr* loss-of-function mutants, express reduced cytotoxicity, but maintained an ability to persist, disseminate and cause abscesses in vivo.

We found no significant enrichments of protein-altering variants among D-class variants, but we observed a significant excess of six protein-altering C-class variants in *pbp2* which encodes a penicillin binding protein involved in cell wall synthesis (19.0-fold enrichment, $p=10^{-6.0}$, ***Figure 2—figure supplement 1A***). Pbp2 is an important target of β-lactam antibiotics (***Łeski and Tomasz, 2005***), revealing adaption – potentially in response to antibiotic treatment – in the nose populations of some patients.

Genes modulated by virulence regulators and antimicrobial peptides show adaptive evolution between colonizing and infecting bacteria

To improve the sensitivity to identify adaptive evolution associated with infection, we developed a gene set enrichment analysis (GSEA) approach in which we tested for enrichments of protein-altering B-class variants among groups of genes. GSEA allowed us to detect signatures of adaptive evolution in groups of related genes that were not apparent when interrogating individual genes.

We grouped genes in two different ways: by gene ontology and by expression pathway. First, we obtained a gene ontology for the reference genome from BioCyc (**Caspi et al., 2016**), which classifies genes into biological processes, cellular components and molecular functions. There were 552 unique gene ontology groupings of two or more genes. We tested for an enrichment among genes belonging to the ontology, compared to the rest of the genes.

Second, we obtained 248 unique expression pathways from the SAMMD database of transcriptional studies (**Nagarajan and Elasri, 2007**). For each expression pathway, genes were classified as up-regulated, down-regulated or not differentially regulated in response to experimentally manipulated growth conditions or expression of a regulatory gene. For each expression pathway, we tested for an enrichment in genes that were up- or down-regulated compared to genes not differentially regulated.

The most significant enrichment for protein-altering B-class variants between nose and infection sites occurred in the group of genes down-regulated by the cationic antimicrobial peptide (CAMP) ovispirin-1 ($p=10^{-7.8}$), with a similar enrichment in genes down-regulated by temporin L exposure ($p=10^{-6.9}$, **Figure 2C**). Like human CAMPs, the animal-derived ovispirin and temporin compounds inhibit epithelial infections by killing phagocytosed bacteria and mediating inflammatory responses (**Pietiäinen et al., 2009**). In response to inhibitory levels of ovispirin and temporin, *agr*, surface-expressed adhesins and secreted toxins are all down-regulated. Collectively, down-regulated genes showed 2.7-fold and 2.8-fold enrichments of adaptive evolution, respectively. Conversely, genes up-regulated in response to CAMPs, including the *vraSR* and *vraDE* cell-wall operons and stress response genes (**Pietiäinen et al., 2009**), exhibited 0.4-fold and 0.7-fold enrichments (i.e. depletions), respectively (**Table 3**). Thus, expression of the genes undergoing adaptive evolution is strongly inhibited in vitro by host-derived antimicrobial peptides.

Genes belonging to the cell wall ontology showed the second most significant enrichment for adaptive evolution ($p=10^{-7.0}$). Genes contributing to this 5.0-fold enrichment included the immunoglobulin-binding *S. aureus* Protein A (*spa*), the serine rich adhesin for platelets (*sasA*), clumping factors A and B (*clfA, clfB*), fibronectin binding protein A (*fnbA*) and bone sialic acid binding protein (*bbp*). The latter four genes contributed to another statistically significant 6.4-fold enrichment of adaptive protein evolution in the cell adhesion ontology ($p=10^{-6.5}$, **Figure 3**). Therefore, there is a general enrichment of surface-expressed host-binding antigens undergoing adaptive evolution.

The *rsp* regulon showed the most significant enrichment among gene sets defined by response to individual bacterial regulators ($p=10^{-6.4}$). Genes down-regulated by *rsp* in exponential phase (**Lei et al., 2011**), including surface antigens and the urease operon, exhibited a 3.6-fold enrichment for adaptive evolution, while up-regulated genes showed 0.6-fold enrichment. So whereas *rsp* loss-of-function mutants were rare per se, genes up-regulated in such mutants were hotspots of within-patient adaptation in infected patients. Since expression is a prerequisite for adaptive protein evolution, this implies there are alternative routes by which genes down-regulated by intact *rsp* can be expressed and thereby play an important role within patients other than direct inactivation of *rsp*.

Loss-of-function in *agr* mutants represent one alternative route, since they exhibit similar phenotypes to *rsp* mutants, with reduced cytotoxicity and increased surface antigen expression, albeit reduced ability to form abscesses (**Das et al., 2016**). We found significant enrichments of genes regulated by *agrA* in two different backgrounds ($p<10^{-4.5}$) and by *sarA* ($p=10^{-4.6}$), underlining the influence of adaptive evolution on both secreted and surface-expressed proteins during infection. We found that expression of genes enriched for protein-altering substitutions was also altered in strains possessing reduced susceptibility to vancomycin, although not in a consistent direction across strains ($p<10^{-4.7}$), and to pine-oil disinfectant ($p=10^{-4.4}$), suggesting such genes may be generally involved in response to harsh environments. All significant signals of enrichment produced neutrality indices exceeding one, consistent with adaptive evolution (**Supplementary file 3**).

Several genes contributed to multiple evolutionary signals, particularly cell-wall anchored proteins involved in adhesion, the infection process and immune evasion (**Foster et al., 2013**), including *fnbA*, *clfA*, *clfB*, *sasA* and *spa*. These multifactorial, partially overlapping signals suggest a large target for selection in adapting to the within-patient environment (**Figure 3**). The fact that we observed no comparable significant enrichments in C-class and D-class protein-altering variants (**Figure 2—figure supplement 1**) indicates that these evolutionary patterns are associated specifically with the infection process.

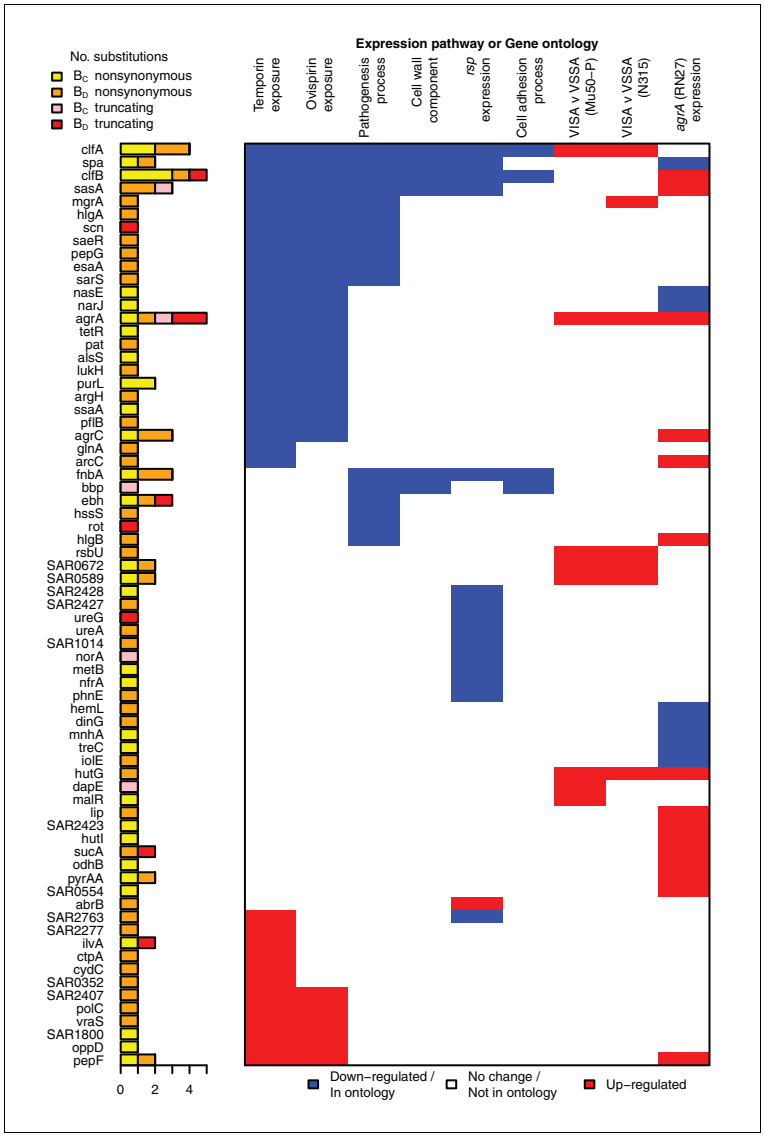

Figure 3. All genes contributing to the pathways and ontologies most significantly enriched for protein-altering substitutions between nose-colonizing and infection-causing bacteria. The pathogenesis ontology, in which significant enrichments were observed in infection-causing but not nose-colonizing bacteria, is shown for comparison. Every gene with at least one substitution between nose-colonizing and infection-causing bacteria and which was up- (red) or down- regulated (blue) in one of the pathways or a member of one of the ontologies (blue) is shown. To the left, the number of altering (yellow/orange) and truncating (pink/red) B-class variants is shown, broken down by the population in which the mutant allele was found: nose (B_C; yellow/pink) or infection site (B_D; orange/red).

DOI: https://doi.org/10.7554/eLife.30637.011

The following figure supplements are available for figure 3:

Figure supplement 1. Genes enriched for substitutions between nose-colonizing and infection-causing bacteria within patients are not the most rapidly evolving at the species level.

DOI: https://doi.org/10.7554/eLife.30637.012

Figure supplement 2. Gene set enrichment analysis is robust to species-level differences in d_N/d_S between genes.

DOI: https://doi.org/10.7554/eLife.30637.013

Adaptive evolution in pathogenesis genes is found only in infecting bacteria

Having identified adaptive evolution differentiating nose-colonizing and infection-causing bacteria, we next asked whether the mutant alleles were preferentially found in the nose or infection site. We used 1149 sequences from other patients or carriers (Reference Panel II) to reconstruct the genotype of the MRCA of nose-colonizing and infecting bacteriain each patient, respectively. This allowed us to sub-classify B-class variants by whether the mutant allele was found in the nose-colonizing bacteria (B_C-class) or the disease-causing infection site bacteria (B_D-class).

A priori, we had expected the enrichments of adaptive evolution to be driven primarily by mutants occurring in the infection-causing bacteria (B_D-class). One group of genes showed a signal of such an enrichment among B_D-class variants specifically. Genes belonging to the BioCyc pathogenesis ontology were marginally genome-wide significant in B_D-class variants, showing a 3.1-fold enrichment ($p=10^{-4.6}$) and a statistically insignificant 1.7-fold enrichment in B_C-class variants ($p=0.13$). B_D-class mutants driving this differential signal arose in toxins including gamma hemolysin and several regulatory loci implicated in toxicity and virulence regulation: *rot*, *sarS* and *saeR*.

Surprisingly, however, we found that all other significantly enriched gene sets were driven by mutant alleles occurring both in nose-colonizing and infecting bacteria (***Figure 2—figure supplement 2***). This indicates there are common selection pressures in the nose and infection site during the process of infection within patients, leading to convergent evolution across body sites. So while adaptation in pathogenesis genes appears specifically infection-associated, other signals of adaptation in severely infected patients are driven by selection pressures, which might compensate for an altered within-host environment during infection, that are as likely to favor mutants in nose-colonizing bacteria as infecting bacteria.

Signals of adaptation are specific to infected patients and differ from prevailing signatures of selection

Two lines of evidence show that the newly discovered signatures of within-host adaptive evolution, both in infecting and nose-colonizing bacteria, are unique to evolution in infected patients. To test this theory against the alternative explanation that our approach merely detects the most rapidly evolving proteins, we searched for similar signals in alternative settings: evolution within asymptomatic carriers, and species-level evolution between unrelated bacteria.

There was no significant enrichment of protein-altering variants in any gene, ontology or pathway among 235 variants identified from 10 longitudinally sampled asymptomatic nasal carriers (Reference Panel III, ***Figure 2—figure supplement 3***, ***Supplementary file 4***). To address the modest sample size, we performed goodness-of-fit tests, focusing on the signals most significantly enriched in patients. We found significant depletions of protein-altering variants in carriers relative to patients in the *rsp*, *agr* and *sarA* regulons ($p=10^{-4.0}$) and the pathogenesis ontology ($p=10^{-3.2}$, ***Supplementary file 5***).

Nor were the relative rates of non-synonymous to synonymous substitution (d_N/d_S) higher between unrelated *S. aureus* (Reference Panel IV) in the genes that contributed most to the signals associated with adaptation within patients: *agrA*, *agrC clfA*, *clfB*, *fnbA* and *sasA*. Although synonymous diversity was somewhat higher than typical in these genes, the d_N/d_S ratios showed no evidence for excess protein-altering change in these compared to other genes (***Figure 3—figure supplement 1***). Accordingly, incorporating this locus-specific variability of d_N/d_S into the GSEA did not affect the results (***Figure 3—figure supplement 2***). Taken together these lines of evidence show that the ontologies, pathways and genes significantly differentiated between nose-colonizing and infecting bacteria arise in response to selection pressures specifically associated with infected patients, and are not repeated in asymptomatic carriers or species-level evolution.

Discussion

We found that common, life-threatening infections of *S. aureus* are frequently descended from commensal bacteria colonizing the nose. These infections are associated with repeatable patterns of bacterial evolution driven by within-patient mutation and selection. Genes involved in pathogenesis, notably toxins and regulators, showed evidence for adaptation in infecting but not nose-colonizing

bacteria. Surprisingly, other signatures of adaptation occurred in parallel in nose-colonizing and infecting bacteria, affecting genes responding to cationic antimicrobial peptides and the virulence regulators *rsp* and *agr*. Such genes mediate toxin production, abscess formation, immune evasion and bacterial-host binding. Adaptation within both regulator and effector genes reveals that multiple, alternative evolutionary paths are targeted by selection in infected patients.

The signatures of within-patient adaptation that we found differed from prevailing signals of selection at the species level. This discordance means that infection-associated adaptive mutations within patients are rarely transmitted, and argues against a straightforward host-pathogen arms race as the predominant evolutionary force acting within and between patients. Instead, it supports the notion of a life-history trade-off between adaptations favoring colonization and infection distinct from those favoring dissemination and onward transmission (**Shopsin et al., 2010**). Infection of this sort can be characterized in terms of an ecological source-sink model, in which commensal bacteria provide the source for potentially life-threatening infections (**Sokurenko et al., 2006**). Ultimately, the short-term survival advantage gained by infecting bacteria, analogous to the short-term advantage of cancerous cells derived from the host, may lead to the demise of both the host and pathogen, epitomizing a tragedy of the commons (**Rankin et al., 2007**). As such, infection may represent an ever-present risk of mutation among commensal bacteria favored by short-term selection but ultimately incidental or damaging to the bacterial reproductive life cycle.

We did not observe analogous signatures of bacterial adaptation and excess loss-of-function mutations in healthy nose carriers, indicating that risk factors for infections, such as a weakened or over-activated immunological response, comorbidities or medical interventions, may play an important role in creating distinctive selection pressures in infected patients. The effects of such risk factors may be mediated, at least in part, through the selection pressure they exert on commensal bacteria.

The existence of signatures of adaptive substitutions associated with infection raises the possibility of developing new diagnostic techniques and personalizing treatment to the individual patient's commensal bacteria. The ability of genomics to characterize the selective forces driving adaption within the human body in unprecedented detail provides new opportunities to improve experimental models of infection. Ultimately, it may be possible to develop therapies that utilize our new understanding of within-patient evolution to target the root causes of infection from the bacterial perspective.

Materials and methods

Patient sample collection

105 patients with severe *S. aureus* infections for whom the organism could be cultured from both admission screening nasal swab and clinical sample were identified prospectively from the microbiological laboratories of hospitals in Oxford and Brighton, England. This study design builds in robustness to potential confounders by matching infection-causing and nose-colonizing bacteria within the same patients. Clinical samples comprised 55 blood cultures and 50 pus, soft tissue, bone or joint samples. The bacteria sampled and sequenced from one patient ('patient S', P005 in this study) have been previously described (**Das et al., 2016**). Five individuals had both blood and another culture-positive clinical sample; we focus analysis on the blood sample. Nasal swabs were incubated in 5% NaCl broth overnight at 37C, then streaked onto SASelect agar (BioRad) and incubated overnight at 37C. We picked five colonies per sample (12 during the pilot phase involving 9 patients), streaked each onto Columbia blood agar and incubated overnight at 37C for DNA extraction. Clinical samples were handled according to the local laboratory standard operating procedure for pus, sterile site and blood cultures. When bacterial growth was confirmed as *S. aureus*, the primary culture plate was retrieved, and multiple colonies were picked. These were streaked onto Columbia blood agar and incubated overnight at 37C for DNA extraction.

Power calculation for within-patient sample sizes

Sequencing multiple colonies per site allowed us to distinguish genuine genetic substitutions between nose-colonizing and infection-causing bacteria from polymorphic variants restricted to the nose-colonizing or infection-causing bacteria respectively. Following previous studies of within-host

S. aureus evolution that we conducted (**Young et al., 2012**; **Golubchik et al., 2013**) and the initial pilot phase in which we sequenced 12 colonies per site, we chose to continue with five colonies per site as a compromise between power to distinguish within-site polymorphisms from true between-site substitutions and the cost of whole genome sequencing. Under a standard neutral model with low mutation rate, five colonies per sample is expected to confer 91% power to correctly distinguish within-site polymorphisms from between-site substitutions, compared to 0% power with one colony per sample. The power calculation is

$$\begin{aligned}\Pr(0<i<n|0<i) &= \frac{\Pr(0<i<n)}{\Pr(0<i)} = \frac{\int_0^1 \Pr(0<i<n)p(f)\mathrm{d}f}{\int_0^1 \Pr(0<i)p(f)\mathrm{d}f} \\ &= \frac{\int_0^1 (1-f^n-(1-f)^n)f^{-1}\mathrm{d}f}{\int_0^1 (1-(1-f)^n)f^{-1}\mathrm{d}f}\end{aligned}$$

where i and f are the observed sample count and unobserved frequency of a mutant allele in a particular site, n is the sample size, and random sampling is assumed. The frequency distribution of a neutral mutant allele, $p(f)$, is derived in (**Sawyer et al., 1992**). If the population were expanding, the power would be greater because mutant alleles would be biased yet more toward low frequencies.

Reference panels

For comparison to the patient-derived bacteria, we collated previously described samples from other sources to construct four Reference Panels: I. A collection of 131 genomes capturing cross-sectional diversity in the noses of 13 asymptomatic carriers (**Golubchik et al., 2013**), arising from a carriage study of *S. aureus* in Oxfordshire (**Everitt et al., 2014**) (BioProject PRJEB2881). II. A compilation of 95 unrelated samples from the same Oxfordshire carriage study (BioProject accession number PRJEB5225), 145 sequences from a study of within-host evolution of *S. aureus* in three individuals (**Young et al., 2012**) (BioProject PRJEB2862) and 909 sequences from nasal carriage and bloodstream infection used in a study of whole genome sequencing to predict antimicrobial resistance (**Gordon et al., 2014**) (BioProject PRJEB5261). We used these samples to improve our reconstruction of ancestral genotypes in each patient. III. A collection of 237 genomes from longitudinal samples from 10 patients (**Golubchik et al., 2013**; **Gordon et al., 2017**), (BioProject PRJNA380544) arising from the same Oxfordshire carriage study. We used these to compare evolution within patients and asymptomatic carriers. IV. A collection of 16 previously published high-quality closed reference genomes, comprising unrelated isolates mainly of clinical and animal origin: MRSA252 (Genbank accession number BX571856.1), MSSA476 (BX571857.1), COL (CP000046.1), NCTC 8325 (CP000253.1), Mu50 (BA000017.4), N315 (BA000018.3), USA300_FPR3757 (CP000255.1), JH1 (CP000736.1), Newman (AP009351.1), TW20 (FN433596.1), S0385 (AM990992.1), JKD6159 (CP002114.2), RF122 (AJ938182.1), ED133 (CP001996.1), ED98 (CP001781.1), EMRSA15 (HE681097.1) (**Holden et al., 2004**; **Gill et al., 2005**; **Gillaspy, 2006**; **Kuroda et al., 2001**; **Diep et al., 2006**; **Baba et al., 2008**; **Holden et al., 2010**; **Schijffelen et al., 2010**; **Chua et al., 2010**; **Herron-Olson et al., 2007**; **Guinane et al., 2010**; **Lowder et al., 2009**; **Holden et al., 2013**). We used these to contrast species-level evolution to within-patient evolution.

Whole genome sequencing

For each bacterial colony, DNA was extracted from the subcultured plate using a mechanical lysis step (FastPrep; MPBiomedicals, Santa Ana, CA) followed by a commercial kit (QuickGene; Fujifilm, Tokyo, Japan), and sequenced at the Wellcome Trust Centre for Human Genetics, Oxford on the Illumina (San Diego, CA) HiSeq 2000 platform, with paired-end reads 101 base pairs for nine patients in the pilot phase, and 150 bases in the remainder. We sequenced 62 genomes in duplicate, a technical replication rate of 5.1%; no genetic discordancies were detected within duplicates.

Variant calling

We used Velvet (**Zerbino and Birney, 2008**) to assemble reads into contigs de novo, and Stampy (**Lunter and Goodson, 2011**) to map reads against two reference genomes: MRSA252 (**Holden et al., 2004**) and a patient-specific reference comprising the contigs assembled for one colony sampled from each patient's nose. Repetitive regions, defined by BLASTing (**Altschul et al., 1990**) the reference genome against itself, were masked prior to variant calling. To obtain multilocus

sequence types (**Enright et al., 2000**), we used BLAST to find the relevant loci, and looked up the nucleotide sequences in the online database at http://saureus.mlst.net/.

Bases called at each position in the reference and those passing previously described (**Young et al., 2012**; **Golubchik et al., 2013**; **Didelot et al., 2012**) quality filters were used to identify single nucleotide polymorphisms (SNPs) from Stampy-based mapping to MRSA252 and the patient-specific reference genomes. We used Cortex (**Iqbal et al., 2012**) to identify SNPs and short indels. Variants found by Cortex were excluded if they had fewer than ten supporting reads or if the base call was heterozygous at more than 5% of reads.

Where physically clustered variants with the same pattern of presence/absence across genomes were found, these were considered likely to represent a single evolutionary event: tandem repeat mutation or recombination. These were de-duplicated to a single variant to avoid inflating evidence of evolutionary events in these regions.

Variant annotation and phylogenetic classification

Maximum likelihood trees were built to infer bacterial relationships within patients (**Gusfield, 1991**). To prioritize variants for further analysis, they were classified according to their phylogenetic position in the tree: B-class (between nose colonization and infection site), C-class (within nose-colonizing population) and D-class (within infection site population). Variants were cross-classified by their predicted functional effect based on mapping to the reference genome or BLASTing to a reference allele: synonymous, non-synonymous or truncating for protein-coding sequences, or non-coding.

Where variation was found using a patient-specific reference, these variants were annotated by first aligning to MRSA252 using Mauve (**Darling et al., 2004**). If no aligned position in MRSA252 could be found, additional annotated references were used. Where variation was found using Cortex only, the variant was annotated by first locating it by comparing the flanking sequence to MRSA252 and other annotated references using BLAST. MRSA252 orthologs were identified using geneDB (**Logan-Klumpler et al., 2012**) and KEGG (**Kanehisa et al., 2016**).

Reconstructing ancestral genotypes per patient

We constructed a species-level phylogeny for all bacteria sampled from the 105 patients together with Reference Panel II (unrelated asymptomatic nose-colonization isolates and bloodstream infection isolates) using a two-step neighbor-joining and maximum likelihood approach, based on a whole-genome alignment derived from mapping all genomes to MRSA252. We first clustered individuals into seven groups using neighbour-joining (**Saitou and Nei, 1987**), before resolving the relationships within each cluster by building a maximum likelihood tree using RAxML (**Stamatakis, 2014**), assuming a general time reversible (GTR) model. To overcome a limitation in the presence of divergent sequences whereby RAxML fixes a minimum branch length that may be longer than a single substitution event, we fine-tuned the estimates of branch lengths using ClonalFrameML (**Didelot and Wilson, 2015**). We used these subtrees to identify, for each patient, the most closely related 'nearest neighbor' sampled from another patient or carrier. We employed this nearest neighbor as an outgroup, and used the tree to reconstruct the sequence of the MRCA of colonizing and infecting bacteria for each patient using a maximum likelihood method (**Pupko et al., 2000**) in ClonalFrameML (**Didelot and Wilson, 2015**). This in turn allowed us to identify the ancestral (wild type) and derived (mutant) allele for all variants mapping to MRSA252. For variants not mapping to MRSA252, we repeated the Cortex variant calling analysis, this time including the nearest neighbor, and identified the ancestral allele as the one possessed by the nearest neighbor. This approach allowed us to identify ancestral (wild type) versus derived (mutant) alleles for 97% of within-patient variants. We used the reconstructions of the within-patient MRCA sequences and identity of ancestral vs derived alleles to sub-categorize B-class variants into those in which the mutant allele was found in the nose-colonizing population (B_C-class) versus the infection-causing population (B_D-class). 521 (97%) of B-class variants were typeable, and in 281 (54%) of these, the mutant allele was found in the infection site population. This allowed us to test for differential enrichments in these two sub-classes.

Mean pairwise genetic diversity

Separately for the nose site and infection site of each patient, we calculated the mean pairwise diversity π as the mean number of variants differing between each pair of genomes. We compared the distributions of π between patients and Reference Panel II (13 cross-sectionally sampled asymptomatic nose carriers) using a Mann-Whitney-Wilcoxon test.

Calculating d_N/d_S ratio

For assessing the d_N/d_S ratio within patients, we adjusted the ratio of raw counts of total numbers of non-synonymous and synonymous SNPs by the ratio expected under strict neutrality. We estimated that the rate of non-synonymous mutation was 4.9 times higher than that of synonymous mutation in *S. aureus* based on codon usage in MRSA252 and the observed transition:transversion ratio in non-coding SNPs.

The neutrality index

To compare the relative d_N/d_S ratios between two groups of variants we computed a Neutrality Index as R_1/R_2 where R_1 and R_2 were the ratio of counts of non-synonymous to synonymous variants in each group respectively (**McDonald and Kreitman, 1991**; **Rand and Kann, 1996**). We compared B-, C- and D-class variants within patients to C-class patients within Reference Panel I (13 cross-sectionally sampled asymptomatic carriers). A Neutrality Index in excess of one indicates a higher d_N/d_S ratio in the former group. We used Fisher's exact test to evaluate the significance of the differences between the groups.

Gene enrichment analysis

To test for significant enrichment of variants in a particular gene, we employed a Poisson regression in which we modelled the expected numbers of de novo variants across patients in any gene j as $\lambda_0 L_j$ under the null hypothesis of no enrichment, where λ_0 gives the expected number of variants per kilobase and L_j is the length of gene j in kilobases. We compared this to the alternative hypothesis in which the expected number of variants was $\lambda_i L_i$ for gene i, the gene of interest, and $\lambda_1 L_j$ for any other gene j. Using R (**R Core Team, 2015**), we estimated the parameters λ_0, λ_1 and λ_i from the data by maximum likelihood and tested for significance via a likelihood ratio test with one degree of freedom. This procedure assumes no recombination within patients, which was reasonable since we found little evidence of recombination in this study or previously (**Golubchik et al., 2013**), including no within-host genetic incompatibilities, and we removed physically clustered variants associated with possible recombination events. We analyzed all protein-coding genes in MRSA252, testing for an enrichment of variants expected to alter the transcribed protein (both non-synonymous and truncating mutations). These tests were also applied to synonymous mutations and no enrichments were found.

Gene set enrichment analysis

Since the number of genes outweighed the number of variants detected, we had limited power to detect weak to modest enrichments at the individual gene level. Instead we pooled genes using ontologies from the BioCyc MRSA252 database (**Caspi et al., 2016**) and expression pathways from the SAMMD database of transcriptional studies (**Nagarajan and Elasri, 2007**). The BioCyc database comprises ontologies describing biological processes, cellular components and molecular functions. The SAMMD database groups genes up-regulated, down-regulated or not differentially regulated in response to experimentally manipulated growth conditions or isogenic mutations, usually of a regulatory gene. After excluding ontologies or pathways with two groups, one involving a single gene, and combining ontologies or pathways with identical groupings of genes, we conducted 800 GSEAs in addition to the 2650 ontologies comprised of individual loci. The number of groupings of genes was always two for BioCyc (included/excluded from the ontology) and two or three for SAMMD (up-/down-/un-differentially regulated in the experiment). Again we employed a Poisson regression in which we modelled the expected numbers of variants in any gene j as $\lambda_0 L_j$ under the null hypothesis of no enrichment where λ_0 gives the expected number of variants per kilobase and L_j is the length of gene j in kilobases. We compared this to the alternative hypothesis in which the expected number of variants was $\lambda_1 L_j$, $\lambda_2 L_j$ or $\lambda_3 L_j$ for gene j depending on the grouping in the ontology/pathway.

Using R, we estimated the parameters λ_0, λ_1, λ_2 and λ_3 from the data by maximum likelihood and tested for significance via a likelihood ratio test with one or two degrees of freedom, depending on the number of groupings in the ontology/pathway.

GSEA multiple testing correction

To account for the multiplicity of testing, we adjusted the p-value significance thresholds from a nominal α = 0.05 using the weighted Bonferroni method. We weighted the significance thresholds by the relative number of tests in each category: 2650 genes, 552 BioCyc ontologies and 248 SAMMD expression pathways. This avoids overly stringent multiple testing correction in categories with fewer tests (***Roeder and Wasserman, 2009***), for example, the 248 SAMMD expression pathways, owing to other categories with very large numbers of tests, for example, the 2650 genes. This gave adjusted significance thresholds of $10^{-5.2}$ for genes, $10^{-4.5}$ for BioCyc ontologies and $10^{-4.2}$ for SAMMD expression pathways.

Longitudinal evolution in asymptomatic carriers

To test whether the patterns of evolution we observed between colonizing and invading bacteria in severely infected patients were typical or unusual, we analyzed Reference Panel III (a collection of 10 longitudinally sampled asymptomatic carriers). Since natural selection is more efficacious over longer periods of time, the longitudinal sampling of these individuals gave us greater opportunity to detect subtle evolutionary patterns in asymptomatic carriers. We characterized variation in these carriers as in the patients. Given the modest sample size and smaller number of variants detected in these individuals (235), we performed GSEA to test for enrichments only in particular genes, ontologies and pathways that were significantly enriched within patients, requiring less stringent multiple testing correction.

omegaMap analysis

We estimated d_N/d_S ratios between unrelated *S. aureus* to characterize the prevailing patterns of selection at the species level. We used Mauve (***Darling et al., 2004***) to pairwise align 15 reference genomes against MRSA252, that is Reference Panel IV. This allowed us to distinguish orthologs from paralogs in the next step in which we multiply aligned all coding sequences overlapping those in MRSA252 using PAGAN (***Löytynoja et al., 2012***). After removing sequences with premature stop codons, we analyzed each alignment of between two and 16 genes using a modification of omegaMap (***Wilson and McVean, 2006***), assuming all sites were unlinked. We previously showed this assumption, which confers substantial computational efficiency savings, does not adversely affect estimates of selection coefficients (***Wilson et al., 2011***). We estimated variation in d_N/d_S within genes using Monte Carlo Markov chain, running each chain for 10,000 iterations. We assumed exponential prior distributions on the population scaled mutation rate (θ), the transition:transversion ratio (κ) and the d_N/d_S ratio (ω) with means 0.05, 3 and 0.2, respectively. We assumed equal codon frequencies and a mean of 30 contiguous codons sharing the same d_N/d_S ratio. For each gene, we computed the posterior mean d_N/d_S ratio across sites. This allowed us to rank the relative strength of selection across genes in MRSA252, and to account for differences in d_N/d_S, as well as gene length, in the GSEA. We achieved this by modifying the expected number of variants in gene j to be $\lambda_0\omega_j L_j$ under the null hypothesis of no enrichment versus $\lambda_1\omega_j L_j$, $\lambda_2\omega_j L_j$ or $\lambda_3\omega_j L_j$ under the alternative hypothesis depending on the ontology or pathway, where ω_j is the posterior mean d_N/d_S in gene j.

Ethical framework

Ethical approval for linking genetic sequences of *S. aureus* isolates to patient data without individual patient consent in Oxford and Brighton in the U.K. was obtained from Berkshire Ethics Committee (10/H0505/83) and the U.K. Health Research Agency [8-05(e)/2010].

Acknowledgements

We would like to thank Ed Feil, Stephen Leslie, Gil McVean and Richard Moxon for helpful insights and useful discussions. Sequencing reads uploaded to short read archive (SRA) under BioProject PRJNA369475. RNAseq data relating to isolate from P005 (aka 'patient S') previously submitted

under BioProject PRJNA279958. The views expressed in this publication are those of the authors and not necessarily those of the funders. This study was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), a Mérieux Research Grant, the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with Public Health England (PHE) (grant HPRU-2012–10041), and the Health Innovation Challenge Fund (a parallel funding partnership between the Wellcome Trust (grant WT098615/Z/12/Z) and the Department of Health (grant HICF-T5-358)). TEP and DWC are NIHR Senior Investigators. DJW and ZI are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and the Royal Society (Grants 101237/Z/13/Z and 102541/Z/13/Z). BCY is a Research Training Fellow funded by the Wellcome Trust (Grant 101611/Z/13/Z). We acknowledge the support of Wellcome Trust Centre for Human Genetics core funding (Grant 090532/Z/09/Z).

Additional information

Funding

Funder	Grant reference number	Author
Wellcome	Health Innovation Challenge Fund WT098615/Z/12/Z	Derrick W Crook
National Institute for Health Research	Oxford NIHR Biomedical Research Centre	Derrick W Crook Timothy E Peto
Department of Health	Health Innovation Challenge Fund HICF-T5-358	Derrick W Crook
Royal Society	Sir Henry Dale Fellowship 101237/Z/13/Z	Daniel J Wilson
Institut Mérieux	Mérieux Research Grant	Rory Bowden Derrick W Crook Daniel J Wilson
Public Health England	HPRU in Healthcare Associated Infections and Antimicrobial Resistance HPRU-2012-10041	Derrick W Crook Timothy E Peto A Sarah Walker
National Institute for Health Research	HPRU in Healthcare Associated Infections and Antimicrobial Resistance HPRU-2012-10041	Derrick W Crook Timothy E Peto A Sarah Walker
Wellcome	Sir Henry Dale Fellowship 101237/Z/13/Z	Daniel J Wilson
Wellcome	Sir Henry Dale Fellowship 102541/Z/13/Z	Zamin Iqbal
Royal Society	Sir Henry Dale Fellowship 102541/Z/13/Z	Zamin Iqbal
Wellcome	Research Training Fellowship 101611/Z/13/Z	Bernadette C Young
Wellcome	Wellcome Trust Centre for Human Genetics core funding 090532/Z/09/Z	Rory Bowden

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Bernadette C Young, Study design, Sample collection, DNA extraction, Bioinformatics, Analysis, Writing; Chieh-Hsi Wu, Bioinformatics, Analysis, Writing; N Claire Gordon, Sample collection, DNA

extraction; Kevin Cole, Elian Liu, Sanuki Perera, DNA extraction; James R Price, Sample collection; Anna E Sheppard, Jane Charlesworth, Tanya Golubchik, Zamin Iqbal, Bioinformatics; Rory Bowden, Ruth C Massey, Study design, Interpretation; John Paul, Derrick W Crook, Timothy E Peto, A Sarah Walker, Martin J Llewelyn, Study design, Sample collection, Interpretation; David H Wyllie, Study design, Analysis; Daniel J Wilson, Study design, Analysis, Writing

Author ORCIDs
Bernadette C Young, https://orcid.org/0000-0001-6071-6770
Ruth C Massey, https://orcid.org/0000-0002-8154-4039
Daniel J Wilson, https://orcid.org/0000-0002-0940-3311

Ethics
Human subjects: Ethical approval for linking genetic sequences of S. aureus isolates to patient data without individual patient consent in Oxford and Brighton in the U.K. was obtained from Berkshire Ethics Committee (10/H0505/83) and the U.K. Health Research Agency [8-05(e)/2010].

Decision letter and Author response
Decision letter https://doi.org/10.7554/eLife.30637.064
Author response https://doi.org/10.7554/eLife.30637.065

Additional files
Supplementary files
• Supplementary file 1. List of all cultures included in the site, the site of infection (and any known source if bloodstream), number of isolates sequenced from each site, ST or CC by in silico MLST, number of variants found at each site and the mean pair-wise difference comparing isolates.
DOI: https://doi.org/10.7554/eLife.30637.014

• Supplementary file 2. List of all variants found within patients with *S. aureus* infections, location on shared reference (MRSA252), or position and reference genome name and accession number if variant could not be localized on MRSA252. Each variant is described by the alleles found, its location in gene, the predicted effect on gene product and the location of the variant on the phylogenetic tree.
DOI: https://doi.org/10.7554/eLife.30637.015

• Supplementary file 3. Neutrality indices show signals of adaptation among the genes, gene ontologies and expression pathways most significantly enriched for protein-altering B-class variants. Neutrality indices (NIs, *41,42*) were calculated as the odds ratio of the number of protein-altering to synonymous variants among B-class versus C/D-class variants. These tests are less powerful than the Poisson regression likelihood ratio tests used to detect gene or gene set enrichment of protein-altering B-class variants (***Table 3***); we present them to demonstrate that the direction of enrichment was consistent with adaptation (NI > 1). To mitigate the reduced power, we calculated the expected numbers of protein-altering B-class variants from the numbers of protein-altering C/D-class variants, synonymous B-class variants and synonymous C/D-class variants by pooling them across all genes. This was justified by the absence of evidence for within-patient recombination and lack of enrichment signals among synonymous variants and C/D class protein-altering variants. A one-tailed Poisson test in R (***R Core Team, 2015***) was used to test NI > 1 (significant NIs at $p<0.05$ in bold).
DOI: https://doi.org/10.7554/eLife.30637.016

• Supplementary file 4. List of all variants found within long term asymptomatic carriers, location on shared reference (MRSA252), or position and reference genome name and accession number if variant was not localized on MRSA252. Each variant is described by the alleles found, its location in gene and the predicted effect on gene product.
DOI: https://doi.org/10.7554/eLife.30637.017

• Supplementary file 5. For all ontologies showing enrichment in within-patient B_D-class variants, we identified the genes with variants contributing to the signal. We counted the number of protein-altering variants in these genes within patients, and compared to the number in long-term asymptomatic carriers. p-Values calculated using Fisher's exact test. *Variant totals are different for

SAMMD pathways (*rsp, agrA, sarA*) and BioCyc ontologies (cell wall, cell adhesion, pathogenesis) because pathway information is available for a different number of loci in each database.

DOI: https://doi.org/10.7554/eLife.30637.018

• Transparent reporting form

DOI: https://doi.org/10.7554/eLife.30637.019

Major datasets

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Bernadette C Young, Chieh-Hsi Wu, N Claire Gordon, James R Price, Kevin Cole, Elian Liu, Anna E Sheppard, Sanuki Perera, Tanya Golubchik, Zamin Iqbal, Rory Bowden, Ruth C Massey, John Paul, Derrick W Crook, Timothy E Peto, A Sarah Walker, Martin J Llewelyn, David H Wyllie, Daniel J Wilson	2017	Illumina Sequencing Data	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA369475	Publicly available at NCBI BioProject (accession no. PRJNA369475)

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database, license, and accessibility information
Tanya Golubchik, Elizabeth M. Batty, Ruth R. Miller, Helen Farr, Bernadette C. Young, Hanna Larner-Svensson, Rowena Fung, Heather Godwin, Kyle Knox, Antonina Votintseva, Richard G. Everitt, Teresa Street, Madeleine Cule, Camilla L. C. Ip, Xavier Didelot, Timothy E. A. Peto, Rosalind M. Harding, Daniel J. Wilson, Derrick W. Crook, Rory Bowden	2013	Reference Panel I	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2881	Publicly available at NCBI BioProject (accession no. PRJEB2881)

Richard G. Everitt, Xavier Didelot, Elizabeth M. Batty, Ruth R Miller, Kyle Knox, Bernadette C. Young, Rory Bowden, Adam Auton, Antonina Votintseva, Hanna Larner-Svensson, Jane Charlesworth, Tanya Golubchik, Camilla L. C. Ip, Heather Godwin, Rowena Fung, Tim E. A. Peto, A. Sarah Walker, Derrick W. Crook & Daniel J. Wilson	2014	Reference Panel IIa	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB5225	Publicly available at NCBI BioProject (accession no. PRJEB5225)
Bernadette C. Young, Tanya Golubchik, Elizabeth Batty, Rowena Fung, Hanna Larner-Svensson, Antonina Votintseva, Ruth Miller, Heather Goodwin, Kyle Knox, Richared Everitt, Zamin Iqbal, Andrew Rimer, Madeline Cule, Camilla Ip, Xavier Didelot, Rosalind Harding, Peter Donnelly, Timothy E Peto, Derrick W Crook, Rory Bowden, Daniel J Wilson	2012	Reference Panel IIb	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB2862	Publicly available at NCBI BioProject (accession no. PRJEB2862)
N Claire Gordon, JR Price, K Cole, R Everitt, M Morgan, J Finney, AM Kearns, B Pichon, BC Young, DJ Wilson, MJ Llewelyn, J. Paul, TEA. Peto, DW Crooa, AS Walker, T Golubchik	2014	Reference Panel IIc	https://www.ncbi.nlm.nih.gov/bioproject/PRJEB5261	Publicly available at NCBI BioProject (accession no. PRJEB5261)
N Claire Gordon, B Pichon, T Golubchik, DJ Wilson, John Paul, DS Blanc, Kevin Cole, J Collins, N Cortes, M Cubbon, FK Gould, PJ Jenks, M Llewelyn, JQ Nash, JM Orendi, K Paranthaman, J Price, L Senn, HL Thomas, S Wyllie, DW Crook, Timothy Peto, AS Walker, AM Kearns	2017	Reference Panel III	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA380544	Publicly available at NCBI BioProject (accession no. PRJNA380544)
Matthew Holden	2004	Reference Panel IV MRSA252	https://www.ncbi.nlm.nih.gov/nuccore/BX571856.1	Publicly available at NCBI Nucleotide (accession no. BX571856.1)

Matthew Holden	2004	Reference Panel IV MSSA476	https://www.ncbi.nlm.nih.gov/nuccore/BX571857.1	Publicly available at NCBI Nucleotide (accession no. BX571857.1)
SR Gill	2005	Reference Panel IV COL	https://www.ncbi.nlm.nih.gov/nuccore/CP000046.1	Publicly available at NCBI Nucleotide (accession no. CP000046.1)
AF Gillaspy	2006	Reference Panel IV NCTC 8325	https://www.ncbi.nlm.nih.gov/nuccore/CP000253.1	Publicly available at NCBI Nucleotide (accession no. CP000253.1)
M Kuroda	2001	Reference Panel IV Mu50	https://www.ncbi.nlm.nih.gov/nuccore/BA000017.4	Publicly available at NCBI Nucleotide (accession no. BA000017.4)
M Kuroda	2001	Reference Panel IV N315	https://www.ncbi.nlm.nih.gov/nuccore/BA000018.3	Publicly available at NCBI Nucleotide (accession no. BA000018.3)
BA Diep	2006	Reference Panel IV USA300_FPR3757	https://www.ncbi.nlm.nih.gov/nuccore/CP000255.1	Publicly available at NCBI Nucleotide (accession no. CP000255.1)
A Copeland	2007	Reference Panel IV JH1	https://www.ncbi.nlm.nih.gov/nuccore/CP000736.1	Publicly available at NCBI Nucleotide (accession no. CP000736.1)
T Baba	2008	Reference Panel IV Newman	https://www.ncbi.nlm.nih.gov/nuccore/AP009351.1	Publicly available at NCBI Nucleotide (accession no. AP009351.1)
Matthew Holden	2010	Reference Panel IV TW20	https://www.ncbi.nlm.nih.gov/nuccore/FN433596.1	Publicly available at NCBI Nucleotide (accession no. FN433596.1)
MJ Schijffelen	2010	Reference Panel IV S0385	https://www.ncbi.nlm.nih.gov/nuccore/AM990992.1	Publicly available at NCBI Nucleotide (accession no. AM990992.1)
K Chua	2010	Reference Panel IV JKD6159	https://www.ncbi.nlm.nih.gov/nuccore/CP002114.2	Publicly available at NCBI Nucleotide (accession no. CP002114.2)
Herron-Olson	2007	Reference Panel IV RF122	https://www.ncbi.nlm.nih.gov/nuccore/AJ938182.1	Publicly available at NCBI Nucleotide (accession no. AJ938182.1)
CM Guinane	2010	Reference Panel IV ED133	https://www.ncbi.nlm.nih.gov/nuccore/CP001996.1	Publicly available at NCBI Nucleotide (accession no. CP001996.1)
BV Lowder	2009	Reference Panel IV ED98	https://www.ncbi.nlm.nih.gov/nuccore/CP001781.1	Publicly available at NCBI Nucleotide (accession no. CP001781.1)
Matthew Holden	2013	Reference Panel IV EMRSA15	https://www.ncbi.nlm.nih.gov/nuccore/HE681097.1	Publicly available at NCBI Nucleotide (accession no. HE681097.1)

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2, PMID: 2231712

Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. 2008. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *Journal of Bacteriology* **190**:300–310. DOI: https://doi.org/10.1128/JB.01000-07, PMID: 17951380

Brown SP, Cornforth DM, Mideo N. 2012. Evolution of virulence in opportunistic pathogens: generalism, plasticity, and control. *Trends in Microbiology* **20**:336–342. DOI: https://doi.org/10.1016/j.tim.2012.04.005, PMID: 22564248

Casadevall A, Fang FC, Pirofski LA, Fang, LA Pirofski FC. 2011. Microbial virulence as an emergent property: consequences and opportunities. *PLoS Pathogens* **7**:e1002136. DOI: https://doi.org/10.1371/journal.ppat.1002136, PMID: 21814511

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* **44**:D471–D480. DOI: https://doi.org/10.1093/nar/gkv1164, PMID: 26527732

Cassat J, Dunman PM, Murphy E, Projan SJ, Beenken KE, Palm KJ, Yang SJ, Rice KC, Bayles KW, Smeltzer MS. 2006. Transcriptional profiling of a *Staphylococcus aureus* clinical isolate and its isogenic agr and sarA mutants reveals global differences in comparison to the laboratory strain RN6390. *Microbiology* **152**:3075–3090. DOI: https://doi.org/10.1099/mic.0.29033-0, PMID: 17005987

Chen PE, Shapiro BJ, Chen, BJ Shapiro PE. 2015. The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology* **25**:17–24. DOI: https://doi.org/10.1016/j.mib.2015.03.002, PMID: 25835153

Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, Hanage WP, Goldblatt D, Nosten FH, Turner C, Turner P, Bentley SD, Parkhill J. 2014. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genetics* **10**:e1004547. DOI: https://doi.org/10.1371/journal.pgen.1004547, PMID: 25101644

Chua K, Seemann T, Harrison PF, Davies JK, Coutts SJ, Chen H, Haring V, Moore R, Howden BP, Stinear TP. 2010. Complete genome sequence of *Staphylococcus aureus* strain JKD6159, a unique Australian clone of ST93-IV community methicillin-resistant *Staphylococcus aureus*. *Journal of Bacteriology* **192**:5556–5557. DOI: https://doi.org/10.1128/JB.00878-10, PMID: 20729356

Cui L, Lian JQ, Neoh HM, Reyes E, Hiramatsu K. 2005. DNA microarray-based identification of genes associated with glycopeptide resistance in *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy* **49**:3404–3413. DOI: https://doi.org/10.1128/AAC.49.8.3404-3413.2005, PMID: 16048954

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**:1394–1403. DOI: https://doi.org/10.1101/gr.2289704, PMID: 15231754

Das S, Lindemann C, Young BC, Muller J, Österreich B, Ternette N, Winkler AC, Paprotka K, Reinhardt R, Förstner KU, Allen E, Flaxman A, Yamaguchi Y, Rollier CS, van Diemen P, Blättner S, Remmele CW, Selle M, Dittrich M, Müller T, et al. 2016. Natural mutations in a *Staphylococcus aureus* virulence regulator attenuate cytotoxicity but permit bacteremia and abscess formation. *PNAS* **113**:E3101–E3110. DOI: https://doi.org/10.1073/pnas.1520255113, PMID: 27185949

Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, Piazza P, Wilson DJ, Bowden R, Donnelly PJ, Dingle KE, Wilcox M, Walker AS, Crook DW, Peto TE, Harding RM. 2012. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biology* **13**:R118. DOI: https://doi.org/10.1186/gb-2012-13-12-r118, PMID: 23259504

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology* **14**:150–162. DOI: https://doi.org/10.1038/nrmicro.2015.13, PMID: 26806595

Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Computational Biology* **11**:e1004041. DOI: https://doi.org/10.1371/journal.pcbi.1004041, PMID: 25675341

Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, Sensabaugh GF, Perdreau-Remington F. 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *The Lancet* **367**:731–739. DOI: https://doi.org/10.1016/S0140-6736(06)68231-7, PMID: 16517273

Dunman PM, Murphy E, Haney S, Palacios D, Tucker-Kellogg G, Wu S, Brown EL, Zagursky RJ, Shlaes D, Projan SJ. 2001. Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the agr and/or sarA loci. *Journal of Bacteriology* **183**:7341–7353. DOI: https://doi.org/10.1128/JB.183.24.7341-7353.2001, PMID: 11717293

Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, Spencer CCA, Iqbal Z, Clifton DA, Hopkins KL, Woodford N, Smith EG, Ismail N, Llewelyn MJ, Peto TE, Crook DW, McVean G, Walker AS, Wilson DJ. 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* **1**:16041. DOI: https://doi.org/10.1038/nmicrobiol.2016.41, PMID: 27572646

Eldholm V, Norheim G, von der Lippe B, Kinander W, Dahle UR, Caugant DA, Mannsåker T, Mengshoel AT, Dyrhol-Riise AM, Balloux F. 2014. Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biology* **15**:490. DOI: https://doi.org/10.1186/s13059-014-0490-3, PMID: 25418686

Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *Journal of Clinical Microbiology* **38**:1008–1015. PMID: 10698988

Everitt RG, Didelot X, Batty EM, Miller RR, Knox K, Young BC, Bowden R, Auton A, Votintseva A, Larner-Svensson H, Charlesworth J, Golubchik T, Ip CL, Godwin H, Fung R, Peto TE, Walker AS, Crook DW, Wilson DJ. 2014. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nature Communications* **5**:3956. DOI: https://doi.org/10.1038/ncomms4956, PMID: 24853639

Foster TJ, Geoghegan JA, Ganesh VK, Höök M. 2013. Adhesion, invasion and evasion: the many functions of the surface proteins of *Staphylococcus aureus*. *Nature Reviews Microbiology* **12**:49–62. DOI: https://doi.org/10.1038/nrmicro3161

Gao W, Chua K, Davies JK, Newton HJ, Seemann T, Harrison PF, Holmes NE, Rhee HW, Hong JI, Hartland EL, Stinear TP, Howden BP. 2010. Two novel point mutations in clinical Staphylococcus aureus reduce linezolid susceptibility and switch on the stringent response to promote persistent infection. *PLoS Pathogens* **6**: e1000944. DOI: https://doi.org/10.1371/journal.ppat.1000944, PMID: 20548948

GBD 2015 Mortality and Causes of Death Collaborators. 2016. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the global burden of disease study 2015. *Lancet* **388**:1459–1544. DOI: https://doi.org/10.1016/S0140-6736(16)31012-1, PMID: 27733281

Gill SR, Fouts DE, Archer GL, Mongodin EF, Deboy RT, Ravel J, Paulsen IT, Kolonay JF, Brinkac L, Beanan M, Dodson RJ, Daugherty SC, Madupu R, Angiuoli SV, Durkin AS, Haft DH, Vamathevan J, Khouri H, Utterback T, Lee C, et al. 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *Journal of Bacteriology* **187**:2426–2438. DOI: https://doi.org/10.1128/JB.187.7.2426-2438.2005, PMID: 15774886

Gillaspy AF. 2006. The *Staphylococcus aureus* NCTC8325 genome. In: Fischetti V, Novick R, Ferretti J (Eds). *Gram Positive Pathogens*. ASM Press.

Golubchik T, Batty EM, Miller RR, Farr H, Young BC, Larner-Svensson H, Fung R, Godwin H, Knox K, Votintseva A, Everitt RG, Street T, Cule M, Ip CL, Didelot X, Peto TE, Harding RM, Wilson DJ, Crook DW, Bowden R. 2013. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One* **8**:e61319. DOI: https://doi.org/10.1371/journal.pone.0061319, PMID: 23658690

Gordon NC, Pichon B, Golubchik T, Wilson DJ, Paul J, Blanc DS, Cole K, Collins J, Cortes N, Cubbon M, Gould FK, Jenks PJ, Llewelyn M, Nash JQ, Orendi JM, Paranthaman K, Price JR, Senn L, Thomas HL, Wyllie S, et al. 2017. Whole-genome sequencing reveals the contribution of long-term carriers in *staphylococcus aureus* outbreak investigation. *Journal of Clinical Microbiology* **55**:2188–2197. DOI: https://doi.org/10.1128/JCM.00363-17, PMID: 28468851

Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, Kearns AM, Pichon B, Young B, Wilson DJ, Llewelyn MJ, Paul J, Peto TE, Crook DW, Walker AS, Golubchik T. 2014. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *Journal of Clinical Microbiology* **52**:1182–1191. DOI: https://doi.org/10.1128/JCM.03117-13, PMID: 24501024

Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder BV, Cartwright RA, Smyth DS, Smyth CJ, Lindsay JA, Gould KA, Witney A, Hinds J, Bollback JP, Rambaut A, Penadés JR, Fitzgerald JR. 2010. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biology and Evolution* **2**:454–466. DOI: https://doi.org/10.1093/gbe/evq031, PMID: 20624747

Gusfield D. 1991. Efficient algorithms for inferring evolutionary trees. *Networks* **21**:19–28. DOI: https://doi.org/10.1002/net.3230210104

Herron-Olson L, Fitzgerald JR, Musser JM, Kapur V. 2007. Molecular correlates of host specialization in *Staphylococcus aureus*. *PLoS One* **2**:e1120. DOI: https://doi.org/10.1371/journal.pone.0001120, PMID: 17971880

Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, Barron A, Bason N, Bentley SD, Chillingworth C, Chillingworth T, Churcher C, Clark L, Corton C, Cronin A, Doggett J, et al. 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *PNAS* **101**:9786–9791. DOI: https://doi.org/10.1073/pnas.0402521101, PMID: 15213324

Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns AM, Hill RL, Edgeworth J, Gould I, Gant V, Cooke J, et al. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Research* **23**:653–664. DOI: https://doi.org/10.1101/gr.147710.112, PMID: 23299977

Holden MT, Lindsay JA, Corton C, Quail MA, Cockfield JD, Pathak S, Batra R, Parkhill J, Bentley SD, Edgeworth JD. 2010. Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *Journal of Bacteriology* **192**:888–892. DOI: https://doi.org/10.1128/JB.01255-09, PMID: 19948800

Howden BP, McEvoy CR, Allen DL, Chua K, Gao W, Harrison PF, Bell J, Coombs G, Bennett-Wood V, Porter JL, Robins-Browne R, Davies JK, Seemann T, Stinear TP. 2011. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. *PLoS Pathogens* **7**:e1002359. DOI: https://doi.org/10.1371/journal.ppat.1002359, PMID: 22102812

Howden BP, Smith DJ, Mansell A, Johnson PD, Ward PB, Stinear TP, Davies JK. 2008. Different bacterial gene expression patterns and attenuated host immune responses are associated with the evolution of low-level vancomycin resistance during persistent methicillin-resistant *Staphylococcus aureus* bacteraemia. *BMC Microbiology* **8**:39. DOI: https://doi.org/10.1186/1471-2180-8-39, PMID: 18304359

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics* **44**:226–232. DOI: https://doi.org/10.1038/ng.1028, PMID: 22231483

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**:D457–D462. DOI: https://doi.org/10.1093/nar/gkv1070, PMID: 26476454

Kluytmans J, van Belkum A, Verbrugh H. 1997. Nasal carriage of *Staphylococcus aureus*: epidemiology, underlying mechanisms, and associated risks. *Clinical Microbiology Reviews* **10**:505–520. PMID: 9227864

Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi NK, Sawano T, et al. 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *The Lancet* **357**:1225–1240. DOI: https://doi.org/10.1016/S0140-6736(00)04403-2, PMID: 11418146

Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden M, Wilson D, Read TD, et al. 2014. Predicting the virulence of MRSA from its genome sequence. *Genome Research* **24**:839–849. DOI: https://doi.org/10.1101/gr.165415.113, PMID: 24717264

Laabei M, Uhlemann AC, Lowy FD, Austin ED, Yokoyama M, Ouadi K, Feil E, Thorpe HA, Williams B, Perkins M, Peacock SJ, Clarke SR, Dordel J, Holden M, Votintseva AA, Bowden R, Crook DW, Young BC, Wilson DJ, Recker M, et al. 2015. Evolutionary trade-offs underlie the multi-faceted virulence of *staphylococcus aureus*. *PLoS Biology* **13**:e1002229. DOI: https://doi.org/10.1371/journal.pbio.1002229, PMID: 26331877

Lamichhane-Khadka R, Riordan JT, Delgado A, Muthaiyan A, Reynolds TD, Wilkinson BJ, Gustafson JE. 2008. Genetic changes that correlate with the pine-oil disinfectant-reduced susceptibility mechanism of *Staphylococcus aureus*. *Journal of Applied Microbiology* **105**:1973–1981. DOI: https://doi.org/10.1111/j.1365-2672.2008.03956.x, PMID: 19120644

Lees JA, Kremer PH, Manso AS, Croucher NJ, Ferwerda B, Serón MV, Oggioni MR, Parkhill J, Brouwer MC, van der Ende A, van de Beek D, Bentley SD. 2017. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microbial Genomics* **3**:e000103. DOI: https://doi.org/10.1099/mgen.0.000103, PMID: 28348877

Lei MG, Cue D, Roux CM, Dunman PM, Lee CY. 2011. Rsp inhibits attachment and biofilm formation by repressing fnbA in *Staphylococcus aureus* MW2. *Journal of Bacteriology* **193**:5231–5241. DOI: https://doi.org/10.1128/JB.05454-11, PMID: 21804010

Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature Genetics* **46**:82–87. DOI: https://doi.org/10.1038/ng.2848, PMID: 24316980

Lieberman TD, Michel JB, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Skurnik D, Leiby N, LiPuma JJ, Goldberg JB, McAdam AJ, Priebe GP, Kishony R. 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics* **43**:1275–1280. DOI: https://doi.org/10.1038/ng.997, PMID: 22081229

Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, Phan I, Farris C, Mitra S, Ramasamy G, Wang H, Tivey A, Jackson A, Houston R, Parkhill J, Holden M, et al. 2012. GeneDB–an annotation database for pathogens. *Nucleic Acids Research* **40**:D98–D108. DOI: https://doi.org/10.1093/nar/gkr1032, PMID: 22116062

Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nübel U, Fitzgerald JR. 2009. Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *PNAS* **106**:19545–19550. DOI: https://doi.org/10.1073/pnas.0909285106, PMID: 19884497

Löytynoja A, Vilella AJ, Goldman N. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* **28**:1684–1691. DOI: https://doi.org/10.1093/bioinformatics/bts198, PMID: 22531217

Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21**:936–939. DOI: https://doi.org/10.1101/gr.111120.110, PMID: 20980556

Łeski TA, Tomasz A. 2005. Role of penicillin-binding protein 2 (PBP2) in the antibiotic susceptibility and cell wall cross-linking of *Staphylococcus aureus*: evidence for the cooperative functioning of PBP2, PBP4, and PBP2A. *Journal of Bacteriology* **187**:1815–1824. DOI: https://doi.org/10.1128/JB.187.5.1815-1824.2005, PMID: 15716453

Margolis E, Levin BR. 2007. Within-host evolution for the invasiveness of commensal bacteria: an experimental study of bacteremias resulting from *Haemophilus influenzae* nasal carriage. *The Journal of Infectious Diseases* **196**:1068–1075. DOI: https://doi.org/10.1086/520934, PMID: 17763330

Markussen T, Marvig RL, Gómez-Lozano M, Aanæs K, Burleigh AE, Høiby N, Johansen HK, Molin S, Jelsbak L. 2014. Environmental heterogeneity drives within-host diversification and evolution of *Pseudomonas aeruginosa*. *mBio* **5**:e01592-14. DOI: https://doi.org/10.1128/mBio.01592-14, PMID: 25227464

Marvig RL, Johansen HK, Molin S, Jelsbak L. 2013. Genome analysis of a transmissible lineage of *pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS Genetics* **9**: e1003741. DOI: https://doi.org/10.1371/journal.pgen.1003741, PMID: 24039595

Marvig RL, Sommer LM, Molin S, Johansen HK. 2015. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nature Genetics* **47**:57–64. DOI: https://doi.org/10.1038/ng.3148, PMID: 25401299

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652–654. DOI: https://doi.org/10.1038/351652a0, PMID: 1904993

Méthot PO, Alizon S. 2014. What is a pathogen? toward a process view of host-parasite interactions. *Virulence* **5**: 775–785. DOI: https://doi.org/10.4161/21505594.2014.960726, PMID: 25483864

Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids. *Science* **353**:380–382. DOI: https://doi.org/10.1126/science.aaf3951, PMID: 27463672

Moxon ER, Murphy PA. 1978. *Haemophilus influenzae* bacteremia and meningitis resulting from survival of a single organism. *PNAS* **75**:1534–1536. DOI: https://doi.org/10.1073/pnas.75.3.1534, PMID: 306628

Nagarajan V, Elasri MO. 2007. SAMMD: *Staphylococcus aureus* microarray meta-database. *BMC Genomics* **8**: 351. DOI: https://doi.org/10.1186/1471-2164-8-351, PMID: 17910768

Novick RP, Geisinger E. 2008. Quorum sensing in staphylococci. *Annual Review of Genetics* **42**:541–564. DOI: https://doi.org/10.1146/annurev.genet.42.110807.091640, PMID: 18713030

Painter KL, Krishna A, Wigneshweraraj S, Edwards AM. 2014. What role does the quorum-sensing accessory gene regulator system play during *Staphylococcus aureus* bacteremia? *Trends in Microbiology* **22**:676–685. DOI: https://doi.org/10.1016/j.tim.2014.09.002, PMID: 25300477

Pietiäinen M, François P, Hyyryläinen HL, Tangomo M, Sass V, Sahl HG, Schrenzel J, Kontinen VP. 2009. Transcriptome analysis of the responses of *Staphylococcus aureus* to antimicrobial peptides and characterization of the roles of vraDE and vraSR in antimicrobial resistance. *BMC Genomics* **10**:429. DOI: https://doi.org/10.1186/1471-2164-10-429, PMID: 19751498

Pollitt EJ, West SA, Crusz SA, Burton-Chellew MN, Diggle SP. 2014. Cooperation, quorum sensing, and evolution of virulence in *Staphylococcus aureus*. *Infection and Immunity* **82**:1045–1051. DOI: https://doi.org/10.1128/IAI.01216-13, PMID: 24343650

Prajsnar TK, Hamilton R, Garcia-Lara J, McVicker G, Williams A, Boots M, Foster SJ, Renshaw SA. 2012. A privileged intraphagocyte niche is responsible for disseminated infection of *Staphylococcus aureus* in a zebrafish model. *Cellular Microbiology* **14**:1600–1619. DOI: https://doi.org/10.1111/j.1462-5822.2012.01826.x, PMID: 22694745

Pupko T, Pe'er I, Shamir R, Graur D. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution* **17**:890–896. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a026369, PMID: 10833195

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing . https://www.R-project.org/

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Molecular Biology and Evolution* **13**:735–748. DOI: https://doi.org/10.1093/oxfordjournals.molbev.a025634, PMID: 8754210

Rankin DJ, Bargum K, Kokko H. 2007. The tragedy of the commons in evolutionary biology. *Trends in Ecology & Evolution* **22**:643–651. DOI: https://doi.org/10.1016/j.tree.2007.07.009, PMID: 17981363

Robinson DA, Monk AB, Cooper JE, Feil EJ, Enright MC. 2005. Evolutionary genetics of the accessory gene regulator (agr) locus in *Staphylococcus aureus*. *Journal of Bacteriology* **187**:8312–8321. DOI: https://doi.org/10.1128/JB.187.24.8312-8321.2005, PMID: 16321935

Roeder K, Wasserman L. 2009. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24**:398–413. DOI: https://doi.org/10.1214/09-STS289, PMID: 20711421

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406–425. PMID: 3447015

Sakoulas G, Moise PA, Rybak MJ. 2009. Accessory gene regulator dysfunction: an advantage for *Staphylococcus aureus* in health-care settings? *The Journal of Infectious Diseases* **199**:1558–1559. DOI: https://doi.org/10.1086/598607, PMID: 19392634

Sawyer SA, Hartl DL, Sawyer, DL Hartl SA. 1992. Population genetics of polymorphism and divergence. *Genetics* **132**:1161–1176. PMID: 1459433

Schijffelen MJ, Boel CH, van Strijp JA, Fluit AC. 2010. Whole genome analysis of a livestock-associated methicillin-resistant *Staphylococcus aureus* ST398 isolate from a case of human endocarditis. *BMC Genomics* **11**:376. DOI: https://doi.org/10.1186/1471-2164-11-376, PMID: 20546576

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MC, Parkhill J, Falush D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *PNAS* **110**:11923–11927. DOI: https://doi.org/10.1073/pnas.1305559110, PMID: 23818615

Shopsin B, Eaton C, Wasserman GA, Mathema B, Adhikari RP, Agolory S, Altman DR, Holzman RS, Kreiswirth BN, Novick RP. 2010. Mutations in agr do not persist in natural populations of methicillin-resistant

Staphylococcus aureus. *The Journal of Infectious Diseases* **202**:1593–1599. DOI: https://doi.org/10.1086/656915, PMID: 20942648

Smyth DS, Kafer JM, Wasserman GA, Velickovic L, Mathema B, Holzman RS, Knipe TA, Becker K, von Eiff C, Peters G, Chen L, Kreiswirth BN, Novick RP, Shopsin B. 2012. Nasal carriage as a source of *agr*-defective *Staphylococcus aureus* bacteremia. *The Journal of Infectious Diseases* **206**:1168–1177. DOI: https://doi.org/10.1093/infdis/jis483, PMID: 22859823

Sokurenko EV, Gomulkiewicz R, Dykhuizen DE, Gomulkiewicz, DE Dykhuizen R. 2006. Source-sink dynamics of virulence evolution. *Nature Reviews Microbiology* **4**:548–555. DOI: https://doi.org/10.1038/nrmicro1446, PMID: 16778839

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. DOI: https://doi.org/10.1093/bioinformatics/btu033, PMID: 24451623

Stapels DA, Ramyar KX, Bischoff M, von Köckritz-Blickwede M, Milder FJ, Ruyken M, Eisenbeis J, McWhorter WJ, Herrmann M, van Kessel KP, Geisbrecht BV, Rooijakkers SH. 2014. *Staphylococcus aureus* secretes a unique class of neutrophil serine protease inhibitors. *PNAS* **111**:13187–13192. DOI: https://doi.org/10.1073/pnas.1407616111, PMID: 25161283

Traber KE, Lee E, Benson S, Corrigan R, Cantera M, Shopsin B, Novick RP. 2008. agr function in clinical *Staphylococcus aureus* isolates. *Microbiology* **154**:2265–2274. DOI: https://doi.org/10.1099/mic.0.2007/011874-0, PMID: 18667559

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The human microbiome project. *Nature* **449**:804–810. DOI: https://doi.org/10.1038/nature06244, PMID: 17943116

von Eiff C, Becker K, Machka K, Stammer H, Peters G. 2001. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. Study Group. *New England Journal of Medicine* **344**:11–16. DOI: https://doi.org/10.1056/NEJM200101043440102, PMID: 11136954

Wilson DJ, Hernandez RD, Andolfatto P, Przeworski M. 2011. A population genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genetics* **7**:e1002395. DOI: https://doi.org/10.1371/journal.pgen.1002395, PMID: 22144911

Wilson DJ, McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**:1411–1425. DOI: https://doi.org/10.1534/genetics.105.044917, PMID: 16387887

Yang ES, Tan J, Eells S, Rieg G, Tagudar G, Miller LG. 2010. Body site colonization in patients with community-associated methicillin-resistant *Staphylococcus aureus* and other types of S. aureus skin infections. *Clinical Microbiology and Infection* **16**:425–431. DOI: https://doi.org/10.1111/j.1469-0691.2009.02836.x, PMID: 19689469

Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, Iqbal Z, Rimmer AJ, Cule M, Ip CL, Didelot X, Harding RM, Donnelly P, Peto TE, Crook DW, Bowden R, et al. 2012. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *PNAS* **109**:4550–4555. DOI: https://doi.org/10.1073/pnas.1113219109, PMID: 22393007

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18**:821–829. DOI: https://doi.org/10.1101/gr.074492.107, PMID: 18349386